Preliminary Data Overview for the New York City TLC Project

Commission Prepared by **Automatidata**

Project Overview

The Automatidata project, in collaboration with the NYC Taxi and Limousine Commission, aims to develop a regression model to estimate taxi fares. Initial tasks include creating a pandas DataFrame, examining data types, and conducting a preliminary data assessment to guide future analysis and improve the model.

Key Insights

- After reviewing the dataset, I identified key variables such as trip_distance, total_amount, and payment_type as the most relevant for fare estimation.
- Anomalies found: missing values in trip_distance and inconsistencies in fare_amount.
- Irrelevant columns: vendor_id may be dropped or deprioritized.

Details

	trip_distance	total_amount
8476	2.60	1200.29
20312	0.00	450.30
13861	33.92	258.21
12511	0.00	233.74
15474	0.00	211.80
6064	32.72	179.06
16379	25.50	157.06
3582	7.30	152.30
11269	0.00	151.82
9280	33.96	150.30
1928	12.50	137.80
10291	31.95	131.80
6708	0.32	126.00
11608	23.00	123.30
908	26.12	121.56

Image Alt - Analyze outliers

- Certain irregularities were present, such as inconsistent fare amounts and null values in variables like trip_distance. Before analysis, additional cleaning will be required.

Next Steps

Fix the data: Make sure the important information is clean and complete.

Combine or change data: Join or modify columns to get more useful information.

Explore the data: Look at the data to understand how different things are related to the cost.